COWLES COMMISSION DISCUSSION PAPER: STATISTICS NO. 376

The Specification Problem in Regression Analysis

H. S. Houthakker

November 19, 1952

The variables with which econometrics deals are almost without exception averages or aggregates of some kind. If we have an equation stating, for instance, that the demand for meat in the U. S. is a function of the price of meat, the price of all other commodities, the disposable income of consumers and a random variable representing "all other influences," then all these variables could be disaggregated into an almost infinite number of components and various combinations of the latter. Not only is each variety of meat available in many forms and grades, all with different prices, but these prices will also vary between localities, and all these different prices will have different influences on demand which no average can completely represent. Many similar examples could be given.

The problem which arises here, viz. what variables to introduce into an equation or model, is a particular case of the more general specification problem. This may be defined as the problem of specifying the variables and the number and mathematical form of the equations that enter into a model which is to be used for one or more stated purposes on the basis of a given set of observations. It might be thought that this

is a problem in economic theory, but the dependence on observations shows that is not the case. Economic theory only provides a very general super structure (and sometimes not even that), which is useful mainly in telling what variables not to include; for instance, production costs should not occur in a demand equation.

The specification problem arises because the number of observations at our disposal is always limited, whereas the number of parameters we may want to take into account is virtually unlimited. Yet we cannot estimate more parameters than there are observations, and actually because there are errors in our equations we can only estimate a smaller number. How to select this smaller number is the problem to which we draw attention here, without having much in the way of results to offer yet. It appears that its importance has not been recognized so far.

Maximum likelihood is the criterion now commonly in use for finding estimation procedures; the simultaneous equations approach [4] was chiefly inspired by the realization that the assumptions underlying maximum likelihood estimation in the single equation case (i.e., the method of least squares) are in general not fulfilled in complete economic systems. Maximum likelihood only applies, however, if the regression model is fixed independently of the observations; otherwise it can easily be carried ad absurdum. For if we have say twenty observations we can attain a likelihood of one by introducing twenty parameters, and there will usually be no reasons of economic theory why we should not do so.

Another criterion (itself contained in the maximum likelihood approach) which thus has to be discarded, is that of unbiasedness. Suppose our "true" equation is of the form $y = \beta_1 x_1 + \ldots + \beta_k x_k + u$, where u is

an error term, but that the number of observation is $n \leq k$, so that we cannot estimate all parameters. Then it will be necessary either to ignore some variables x_i altogether, which will clearly make estimates of y biased, or to lump a number of x_i together into a new variable x_i^0. The latter procedure will cause no bias if correct weights are used informing the aggregate, but these correct weights are the parameters β_i which are unknown. As a prescription for estimation unbiasedness can therefore not be taken seriously.

We now need another criterion of optimality which is not subject to this dependence on a fixed model. In principle optimality will be related to the purposes the estimates are to serve. These purposes may be more or less limited and it is worth trying to find criteria which will be applicable to a wide class of purposes. More particularly, we may be interested in the value of an endogenous variable for certain sets of values of the exogenous variables in a model. It may well be that the specification will have to be different for each such set of arguments; we cannot yet assert this with certainty. But there is one criterion which merits consideration even though it may not be optimal for any given set of arguments but only in some average sense.

This rule consists in minimizing the generalized sampling variance of all the estimates around their true values; it is related to Aitken's minimum variance approach ([1], [2]) except in that the latter also requires unbiasedness. Let us again consider the linear equation

$$y = \sum_i \beta_i x_i + u \qquad (1)$$

then the specification will involve a separation of the x_i into two classes:

those that are introduced explicitly (either individually or in combination) and those that are left out of account. The retained variables, supposed to be numbered $x_1, \ldots, x_m$, are transformed linearly into another set $z_1, \ldots, z_p$ ($p \leq m$) and the others are lumped together into one variable v. The equation then reads

$$y = \sum_i \gamma_i z_i + \gamma_v v + u \quad . \tag{2}$$

Denoting the covariance of the estimates c_j and c_k around their true values γ_j and γ_k by $m(c_j, c_k)$ the generalized sample variance, which is to be minimized, becomes

$$|M| = \begin{vmatrix} m(c_1, c_1) & \cdots & m(c_1, c_p) & m(c_1, c_v) \\ \vdots & & \vdots & \vdots \\ m(c_p, c_1) & \cdots & m(c_p, c_p) & m(c_p, c_v) \\ m(c_v, c_1) & \cdots & m(c_v, c_p) & m(c_v, c_v) \end{vmatrix} \quad . \tag{3}$$

We remark that the estimate c_v of γ_v in (2) is zero, signifying that v is the aggregate of the discarded variables. Consequently

$$m(c_v, c_v) = \gamma_v^2 \tag{4}$$

and

$$m(c_v, c_i) = \gamma_v E \qquad i \neq v \quad . \tag{5}$$

It can now already be seen why the discarded variables have to be combined into one; if this were not done the variance determinant $|M|$ would vanish identically.

The transformation of the x_i into the z_i and v can be described by

$$\left\{x_1, \ldots, x_m; x_{m+1}, \ldots, x_k\right\} \begin{bmatrix} a_{11} & \cdots & a_{1p} & 0 \\ \vdots & & \vdots & \vdots \\ a_{m1} & \cdots & a_{mp} & 0 \\ \hline 0 & \cdots & 0 & \beta_{m+1} \\ \vdots & & \vdots & \vdots \\ 0 & \cdots & 0 & \beta_k \end{bmatrix} = \left\{z_1, \ldots, z_p; v\right\} \tag{6}$$

or more briefly by

$$\left\{x \vdots x^*\right\} \left[\begin{array}{c:c} A & 0 \\ \hdashline 0 & \beta^* \end{array}\right] = \left\{z \vdots v\right\} \tag{7}$$

and the specification rule now under discussion consists in choosing a partitioned matrix such that $|M|$ is minimized. In practice the submatrix A will consist mainly of zeros and ones, so that the problem is somewhat simplified; the β_i in (1) and the γ_i in (2) will then be largely the same parameters.

Although this approach can be developed in more detail and related to some results on a similar problem in an unpublished paper by J. Durbin, no small sample rule for selecting the specification has been found yet. The rule-of-thumb applied in [3] appears to be justified in large samples only, and even that is still conjectural.

It also seems probable that the optimal specification depends on the set of values from which a prediction is made. For instance, if we want a prediction in a case where one predetermined variable has a value very different from its mean value in the sample of observations, then it may be better to retain this variable even if it would have been discarded by the criterion of minimum variance of the parameters.

We have so far spoken mainly about single equation regression, but the specification problem is perhaps even more important in simultaneous equation system, where a special case of it is the designation of variables as exogenous or endogenous. This whole subject still awaits exploration.

REFERENCES

[1] AITKEN, A. C., "On Least Squares and Linear Combinations of Observations," Proceedings of the Royal Society of Edinburgh, 55 (1935), p. 42.

[2] _____, "On the Estimation of Many Statistical Parameters," ibid., 62 (1948), p. 369.

[3] HOUTHAKKER, H. S. and J. TOBIN, "Estimates of the Free Demand for Rationed Foodstuffs," Economic Journal, 62 (1952), p. 103, esp. Appendix C.

[4] KOOPMANS, T. C. (ed.), Statistical Inference in Dynamic Economic Models, New York, John Wiley & Sons, 1950.